



Regular expression-based learning to extract bodyweight values from clinical notes



Maureen A. Murtaugh^{a,b,*}, Bryan Smith Gibson^{a,b}, Doug Redd^{a,c}, Qing Zeng-Treitler^{a,c}

^a IDEAS Center, Veterans Administration, Salt Lake City Health Care System, Salt Lake City, UT, United States

^b Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States

^c Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, United States

ARTICLE INFO

Article history:

Received 6 November 2014

Accepted 24 February 2015

Available online 5 March 2015

Keywords:

Natural language processing

Bodyweight

Text classification

ABSTRACT

Background: Bodyweight related measures (weight, height, BMI, abdominal circumference) are extremely important for clinical care, research and quality improvement. These and other vitals signs data are frequently missing from structured tables of electronic health records. However they are often recorded as text within clinical notes. In this project we sought to develop and validate a learning algorithm that would extract bodyweight related measures from clinical notes in the Veterans Administration (VA) Electronic Health Record to complement the structured data used in clinical research.

Methods: We developed the Regular Expression Discovery Extractor (REDEX), a supervised learning algorithm that generates regular expressions from a training set. The regular expressions generated by REDEX were then used to extract the numerical values of interest.

Methods: To train the algorithm we created a corpus of 268 outpatient primary care notes that were annotated by two annotators. This annotation served to develop the annotation process and identify terms associated with bodyweight related measures for training the supervised learning algorithm. Snippets from an additional 300 outpatient primary care notes were subsequently annotated independently by two reviewers to complete the training set. Inter-annotator agreement was calculated.

Methods: REDEX was applied to a separate test set of 3561 notes to generate a dataset of weights extracted from text. We estimated the number of unique individuals who would otherwise not have bodyweight related measures recorded in the CDW and the number of additional bodyweight related measures that would be additionally captured.

Results: REDEX's performance was: accuracy = 98.3%, precision = 98.8%, recall = 98.3%, $F = 98.5\%$. In the dataset of weights from 3561 notes, 7.7% of notes contained bodyweight related measures that were not available as structured data. In addition 2 additional bodyweight related measures were identified per individual per year.

Conclusion: Bodyweight related measures are frequently stored as text in clinical notes. A supervised learning algorithm can be used to extract this data. Implications for clinical care, epidemiology, and quality improvement efforts are discussed.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The use of Electronic Health Record (EHR) data in conjunction with data extraction and categorization tools (e.g. clinical phenotyping), holds great potential to improve clinical practice [6,10] and clinical epidemiology [2]. However, challenges related to data completeness and data quality need to be addressed to maximize

the effectiveness of these efforts. For example bodyweight related measures (weight, height, abdominal circumference), are needed when clinicians calculate medication dosages based on body surface area (BSA) [11], or use body mass index (BMI) to estimate risk of cardiovascular disease, diabetes or cancer (Institute). Similarly, epidemiologists rely on bodyweight measures when determining novel associations such as the recently reported association between bodyweight and mortality due to influenza and pneumonia [4].

Despite the critical importance of bodyweight data for clinical care and research, several evaluations have pointed out that these

* Corresponding author at: Division of Epidemiology, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84132, United States. Fax: +1 (801) 581 3623.

E-mail address: Maureen.Murtaugh@hsc.utah.edu (M.A. Murtaugh).

data are frequently unavailable as structured data. For example researchers at the group health cooperative, testing the ability to use EHR data to calculate cardiac risk, found that among the records of 122,270 individuals, 11.5% were missing data for either height weight or both [5]. Similarly, Das et al. reported that among 1.8 Million Veterans who received outpatient care at VA facilities in the year 2000, 50.4% had no height or weight recorded as structured data [3]. More recently, Littman et al. reported that 32.8% of records of 173,127 veterans in the northwestern US were missing structured data for weight or height [7]. Since anecdotal reports suggested that in many cases individuals' heights and weights were measured during these visits, but the data was recorded as

text in the clinical note, our research team felt that this was an important use case for information extraction.

In this project we sought to develop and validate a learning algorithm that would extract bodyweight related measures (weight, height, BMI, abdominal circumference) recorded in clinical notes from the VA's electronic Health record. We were motivated to explore this as an example of the potential to supplement structured data with data stored in text in order to fill in gaps in repeatedly measured clinical data. Our first aim was to determine how well we could capture weight, height, BMI and/or abdominal girth from outpatient notes. Our second aim was to determine the proportion of the data in the notes that was unique data.

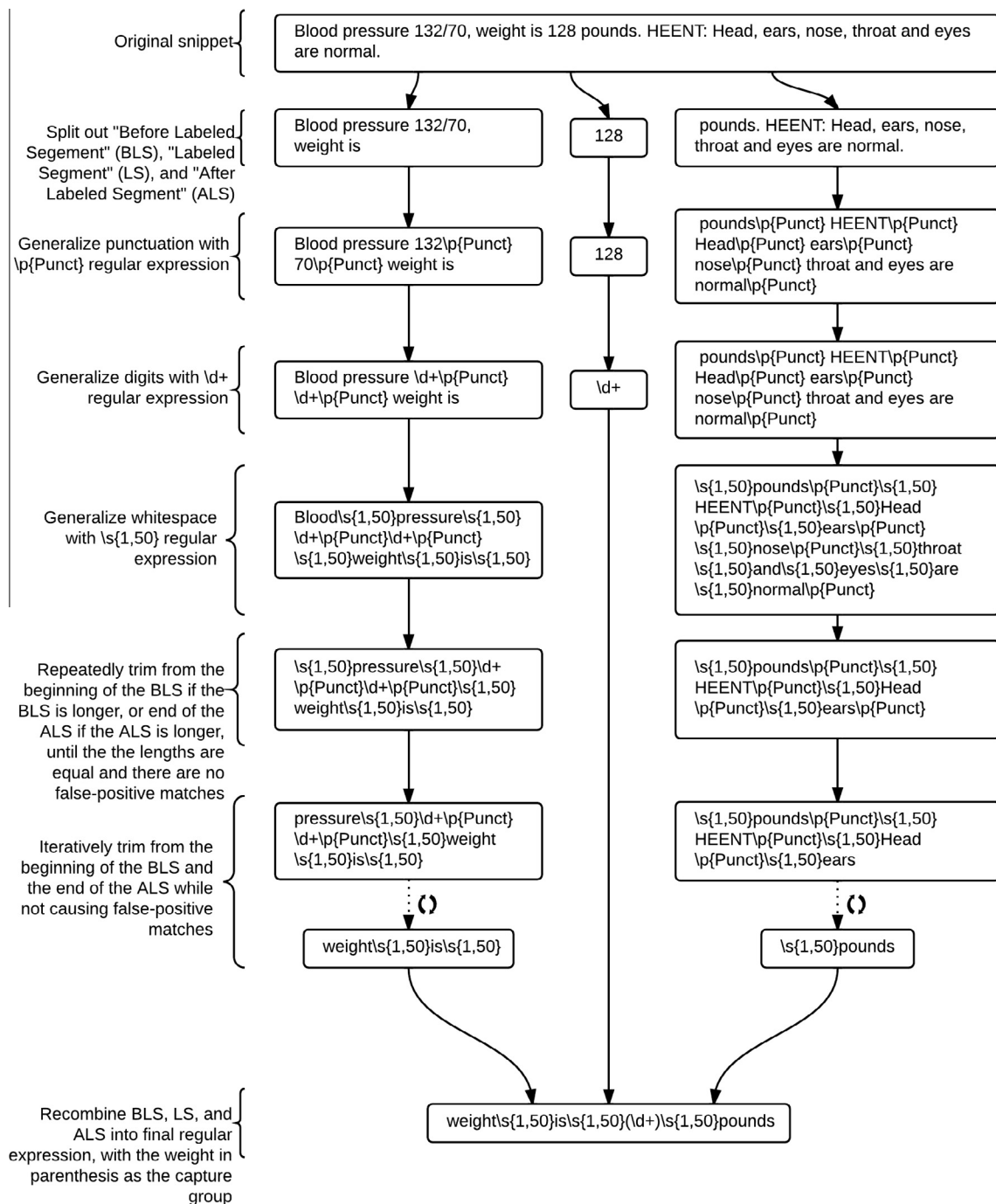


Fig. 1. Example of the creation of a standardized regular expression by REDEX.

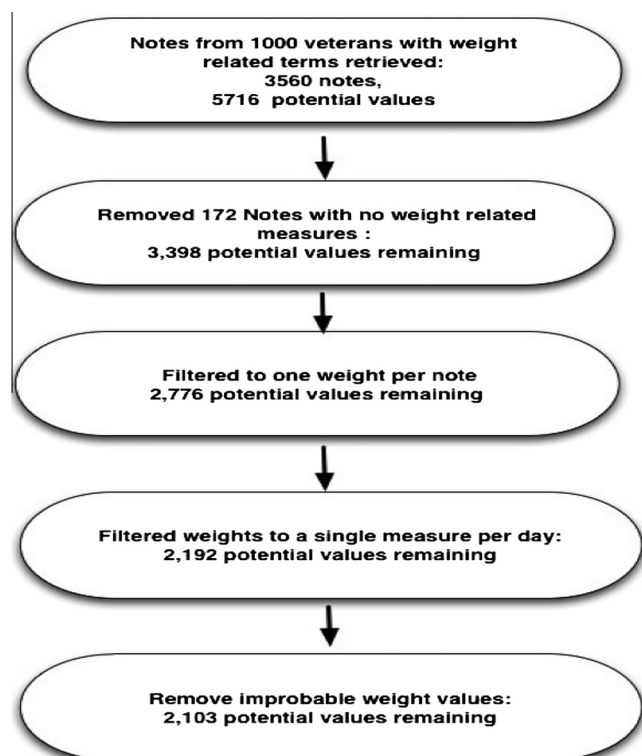


Fig. 2. Presents the data cleaning procedures used to ensure that we extracted only weight.

2. Background

2.1. Veterans Health Administration Informatics and Computing Infrastructure

The Veterans Health Administration was at the forefront of the development of Electronic Health Records and implemented its independently developed EHR, the Veterans Health Information Systems and Technology Architecture (Vista), in 1996. Therefore the Veterans Administration (VA) now has extensive longitudinal records on millions of Veterans.

Recognizing the opportunities for research using this aggregated data, the VA Health Services Research and Development (HSR&D) office funded the Veterans Informatics and Computing Infrastructure (VINCI) [12], a service-level collaboration between the Office of Information and Technology (OI&T) and the Office of Research and Development (OR&D). Designed to serve the data and Information technology needs of the VA research community, VINCI provides secure, centralized access to VA data resources in a high-performance computing environment. VINCI's mission is to provide researchers with an environment for efficient, secure analysis of patient level data, and to provide tools and coordination for research in basic and applied medical informatics.

As of FY 2013, VINCI provides access to structured and unstructured electronic medical data on 17,543,172 unique Veterans. The document corpus consists of 2,096,957,070 clinical documents from providers. The dataset also includes 1,611,284,360 diagnostic codes (ICD9), data on 1,654,598,048 pharmacy prescriptions, and 5,856,426,293 lab tests (both orders and results). Many other types of administrative and clinical data are also available.

2.2. Regular expression based learning

Regular expression-based learning has been an active area of research in computer science and to a lesser degree in biomedical informatics. Some learning algorithms require “seed” expressions, while others are designed to be totally automated. In the biomedical informatics domain, there is no completely automated learning algorithm for generating regular expressions that can be used to extract specific types of numerical values.

The goal of this project was to develop and test a Regular Expression Discovery extraction algorithm (REDEx, Fig. 1) that would address problems in typical regular expression based information extraction. First, the extraction of numerical values from clinical notes is typically performed using manually created regular expressions. This is a laborious process and its accuracy is dependent on the developers' expertise. Maintenance and extensions can also be particularly challenging as there is no standard method to document regular expressions and the patterns that match them. Finally, clinical domain experts who know best what they want to extract are generally not regular expressions experts, thus requiring additional labor to create libraries of regular expressions for more complex domains.

3. Materials and methods

We first retrieved and annotated a set of relevant outpatient notes as a reference standard. We then developed an NLP module for bodyweight related information extraction using the REDEx algorithm. Finally we applied the NLP module to a separate dataset to estimate the value of adding text data to structured data.

3.1. Retrieval of relevant notes

In order to collect the notes for annotation and use in the training of the classifier, we used Voogo [8,13], a search engine developed by our research group specifically to query VINCI data. It supports both free text and structured data searches and provides document, patient, and population-level results. Query results can be lists of patients and related documents, or summary reports that include the geographical distribution, age distribution, living/deceased status, gender, and prescribed VA medications for the veterans about whom the documents were written.

Table 1 provides the list of terms we used to retrieve notes that might contain bodyweight related measures. This list of terms was developed iteratively. We started with a preliminary set of search terms. Two annotators reviewed 268 notes that contained the initial terms. These terms were subsequently modified in order to ensure we were retrieving relevant notes (e.g. the initial search terms included abdominal, and abd* which resulted in many notes that referenced physical exam of the abdomen but were not relevant to our purpose). The annotators assessed the coverage of the initial terms in these notes. The final list of keywords was used to retrieve a second set of text snippets for annotation from 300

Table 1
Search terms used in Voogo to retrieve notes for training set.

Concept	Terms
Bodyweight	wt, weight, wgt, #, lb, kg
Height	height, ht, hgt
BMI	bmi, ibw, ibmi,
Abdominal	abdominal circumference, circumference, girth, waist
Circumference	circumference, whr, waist to hip ratio

notes and the test set of 3561 notes used to create the database (test set) of weights.

3.2. Annotation

We used an annotation tool developed at the National Libraries of Medicine (NLM) called VTT (Visual Tagging tool). This tool allows users to visually tag specific terms of interest in the clinical text that are instantiations of concepts of interest, the output of this tool includes unique identifiers for the notes containing the concept of interest, and the presence of specific tagged terms as coded data. Two researchers (BG and MM) independently annotated text snippets from 300 notes for the presence of the body weight-related measures of interest (weight, height, BMI, abdominal circumference). Text snippets are chunks of text of a limited length that may cross sentences, phrases or boundaries. The text snippets used in this study included the term of interest with a span of 20 words before and after.

Inter-annotator agreement between the two annotators varied based on the measure of interest but overall was excellent: the Kappa value for the weights extracted from the 968 snippets (extracted from 568 notes) was 99.54% for weight (kg, lbs, BMI, height, inches, cm, and feet) and 100% for abdominal circumference (included waist, cm inches, girth, $n = 22$). These annotated snippets were used to create the Regular Expression Discovery extraction algorithm (REDEX).

3.3. Regular Expression Discovery extraction algorithm (REDEX)

The REDEX algorithm builds upon our prior work [1] and contains the following main steps.

1. Each annotated snippet is split into parts: the labeled segment (LS, which is the text annotated by the researchers), before labeled segment (BLS, the text in the snippet preceding the LS), and after labeled segment (ALS, the text in the snippet following the LS).
2. The LS, BLS, and ALS are then converted into generalized regular expressions by first replacing all punctuation, digits, and whitespace with generalized expressions matching any punctuation, digits, or whitespace (e.g. “\p{Punct}”, “\d+”, or “\s{1,50}Br” respectively).
3. Each BLS-LS-ALS triplet is then progressively generalized by successively trimming from the front of the BLS and the end of the ALS until one or more false matches occurs.
4. Redundant triplets are then removed, and each remaining BLS-LS-ALS triplet is converted to a single regular expression, using the LS piece as the regular expression capture group.

The resultant set of regular expressions is then used as the “model” for subsequent extractions. Examples of snippets containing possible expression of weight are found in [Appendix 2](#).

3.4. Final notes selection and data cleaning

We applied the REDEX algorithm to 3560 notes from 1000 Veterans selected at random spanning a period of October 1, 2011 through September 30, 2013 to identify 5716 probable weight values ([Fig. 2](#)). In this dataset of weight values extracted using REDEX, we identified 172 (of 5712) values that were obviously not weights. These appeared to be blood pressure (one value/another value), time (e.g. 1:00 PM), or a range of two values (e.g. 99–101, [Appendix 1](#)). We then filtered the values to include one per note. Next we filtered weights to include only one measurement per day, using the first measurement of the day when

there were multiples (yielding 2192 values). Lastly, we cleaned values to remove weight values <75 and >600 lbs.

3.5. Confirmation of values from text and notes

In order to determine the concordance between bodyweight related values from text and the bodyweight values in the structured data, we calculated the correlation (Pearson's R) between pairs of values. The pairs included one that was taken from notes within one day of its' pair that was recorded as structured field data. We used SAS (Version 9.3) to calculate this correlation.

3.6. Estimation of the proportion of measure that were unique

To estimate the proportion of measures found by the REDEX algorithm that were unique (not duplicating data stored in the structured data tables) we identified weights that were extracted from the text for which there was no related structured data within 1 day of the date of the note.

4. Results

4.1. Accuracy of extraction

The accuracy of REDEX was measured using annotations (actual values) from text snippets. A classification for a snippet was considered a true positive (TP) if the algorithm extracted a value that matched an actual value. A prediction was considered a false positive (FP) if the extractor yielded a value that did not match the actual value, or there was no actual value for the snippet. It was considered a false negative (FN) if the extractor did not predict a value when there was an actual value. It was considered a true negative (TN) when there was no predicted value and there was also no actual value for a snippet. [Table 2](#) presents the confusion matrix for REDEX evaluated using 968 snippets extracted from the 568 manually annotated notes. Evaluation was performed using 10-fold cross validation. Accuracy of the extractor was 98.3%, precision was 98.8%, recall was 98.3%, specificity was 98.1%, and *F*-score was 98.5%.

4.2. Confirmation of weights extracted from text

We used REDEX to select weight values from notes of Veterans who also had a weight in the structured field to confirm the reliability of the values extracted from text. The agreement of weights extracted from outpatient text notes with those extracted from the structured weight field within one day of each other was high (Pearson Correlation, $r = 0.95$).

4.3. Number of unique measures captured by algorithm

The proportion of weight values found by REDEX that were unique measures was 162 of 2103 values, or 7.7%.

5. Discussion

In this paper we demonstrated that REDEX can be used to accurately find unique bodyweight related values in text. Compared to

Table 2
Confusion matrix for weight extractor applied to 968 snippets from 568 notes.

	Actual	
Predicted	584 ^{TP} 10 ^{FN}	7 ^{FP} 367 ^{TN}

manually developing a set of regular expressions for the particular clinical task at hand, the automated learning algorithm provides a more generalizable solution. Extraction of these values had significant impact on both the numbers of unique individuals with bodyweight related measures and the numbers of measures for each individual. These findings suggest that this method could be used more generally for both clinical and research cohort identification to reduce missing data and to improve estimation of longitudinal trajectories of repeated measures within individuals.

Our findings echo other studies that have pointed to deficiencies in the availability of data in structured fields from electronic health records for bodyweight related measures. For example Green et al. reported that 11.5% of 122,270 patients were missing data in the EHR necessary to calculate BMI. Similarly Rose et al. reported that among 79,947 patients served by a large primary care network, 39% (range 6–66%) did not have either height or weight recorded to allow for calculation of Body mass Index (BMI) [9]. The advantage of this study is that it takes the next step by developing an algorithm to extract data from clinical notes and address this missing data problem inherent in clinical care databases.

The impact of additional information from text notes is in part related to the magnitude of the Veterans Health Care system. For example, if this method were applied to the full VINCI database that represents 17 Million unique individuals and identified 7.7% new data points, it would identify 1,309,000 unique values. These values would provide data for individuals who would otherwise not have a weight related measure available for quantitative analysis within a 2-year time-period. Additionally, the methods would identify an average of two additional measures per person per year.

5.1. Strengths/limitations

This method could be applied to other variables in other corpus for a variety of clinical and research domains. For example, the method can be trained to recognize tumor margins or ejection fraction. Limitations include that the algorithm was trained to expressions appearing in the electronic health record at the VA. Importantly, REDEX allows flexibility in that it can be retrained if conventions for expressing the value of interest changes or are different in a different region or system. Methods that do not allow any false positives could lead to over fitting. In this particular use case though, we did not observe signs of over fitting in the expressions that were generated.

5.2. Future work

The extraction of these values enables us to evaluate a number of clinical interventions related to obesity and chronic disease. Future work can address the important question of whether the capture of this data would change the classification of individuals regarding their weight/obesity status and evaluate the impact on assessment of clinical outcomes related to changes in body weight or other vital signs. We are experimenting with permitting a small number of false positives in training, to avoid over-fitting and to tolerate a certain level annotation errors. Additionally, we will be able to demonstrate how the capture of data using NLP changes the estimation of trajectories in important vital sign data. Further, application of this method can be used to reduce potential sampling bias related to missing or mistimed clinical data.

6. Conclusion

A regular expression based learning algorithm to extract measures of interest in clinical text notes performed with high accuracy. The method extracts unique values (not stored elsewhere on the clinical database) for Veterans who did have existing values. The method also identified additional values for Veterans who had some existing values. We believe this approach offers value for improving clinical data completeness and quality for both research and evaluating clinical care.

Acknowledgments

Funding for this project came from the following sources:

HIR 08-374: VA Health Services Research and Development Consortium for Healthcare Informatics Research (CHIR).

HIR 08-204: Veterans Affairs Health Services Research & Development, VA Informatics and Computing Infrastructure (VINCI) aka Center for Scientific Computing.

CRE 12-315: Veterans Affairs Health Services Research & Development, CREATE: A VHA NLP Software Ecosystem for Collaborative Development and Integration.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jbi.2015.02.009>.

References

- [1] Bui DD, Zeng-Treitler Q. Learning regular expressions for clinical text classification. *J Am Med Inform Assoc* 2014;21:850–7.
- [2] Chute CG. Invited commentary: observational research in the age of the electronic health record. *Am J Epidemiol* 2014.
- [3] Das SR, Kinsinger LS, Yancy Jr W, Wang A, Ciesco E, Burdick M, et al. Obesity prevalence among veterans at Veterans Affairs medical facilities. *Am J Prev Med* 2005;28:291–4.
- [4] Fisher-Hoch SP, Mathews CE, McCormick JB. Obesity, diabetes and pneumonia: the menacing interface of non-communicable and infectious diseases. *Trop Med Int Heal* 2013;18:1510–9.
- [5] Green BB, Anderson ML, Cook AJ, Catz S, Fishman PA, McClure JB, et al. Using body mass index data in the electronic health record to calculate cardiovascular risk. *Am J Prev Med* 2012;42:342–7.
- [6] Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol* 2013.
- [7] Littman A, Boyko EJ, McDonnell M, Fihn S. Evaluation of a weight management program for veterans. *Prev Chron Dis* 2012;9:110267.
- [8] Redd D, Rindfleisch T, Nebeker J, Zeng-Treitler Q. Improve retrieval performance on clinical notes: a comparison of four methods. *IEEE*; 2013. p. 2389–97.
- [9] Rose SA, Turchin A, Grant RW, Meigs JB. Documentation of body mass index and control of associated risk factors in a large primary care network. *BMC Heal Serv Res* 2009;9:236.
- [10] Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. *AMIA Summits Transl Sci Proc* 2013;2013:249–53.
- [11] Tipton P, Aigner M, Finto D, Haislet J, Pehl L, Sanford P, et al. Patient safety: consider the accuracy of height and weight measurements. *Nursing* 2012;42:50–2.
- [12] US Department of Veterans Affairs. VA Informatics and Computing Infrastructure (VINCI) Washington DC: US Department of Veterans Affairs; 2013. <http://www.hsrd.research.va.gov/for_researchers/vinci/2013>.
- [13] Zeng QT, Redd D, Rindfleisch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. *American Medical Informatics Association*; 2012. 1050.